Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 35 (2016) 241 – 248

Procedia Economics and Finance

www.elsevier.com/locate/procedia

7th International Economics & Business Management Conference, 5th & 6th October 2015

Panel Data Analysis for Sabah Construction Industries: Choosing the Best Model

Anwar Fitrianto[a,b,c]*, Nur Farhanah Kahal Musakkal[a]

[a]*Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Malaysia*
[b]*Laboratory of Computational Statistics and Operations Research, Institute for Mathematical Research, Univeriti Putra Malaysia*
[c]*Department of Statistics, Faculty of Mathematics and Natural Resources, Bogor Agricultural University, Indonesia*

Abstract

Analysis of panel data by using statistical models is rapidly growing. It is sometime tough for the novice users of panel data to make an informed choice of what estimators best suit their research questions. This paper is meant to find best model among few types of models such as panel data models and ordinary least squares (OLS) regression for Sabah construction industries. The best model will be chosen based on lowest Root Mean Square Errors (RMSE). The purpose of comparing between models is to find the most efficient model which will be useful for prediction. After analyzing the data using SAS software, it was found that two-way fixed effect panel data model provide the lowest RMSE for the Sabah construction industries.

Peer-reviewed under responsibility of Universiti Tenaga Nasional

Keywords: Panel data, Regression, Time series, Random effects, Fixed effects, Sabah

1. Introduction

Econometric analysis of panel data has been started some time ago, such as what has been done by Balestra and Nerlove (1966) and Hoch (1962). Panel data analysis can benefit to industries because it provides information from dataset which behavior of cross sectional is observed across time. The pooling of cross section and time series data

* Corresponding author. Tel.: +6038976-7025
E-mail address: anwarstat@gmail.com

Peer-reviewed under responsibility of Universiti Tenaga Nasional
doi:10.1016/S2212-5671(16)00030-7

have been increasingly important and popular way to determine economic relationship. In panel data, each series yields information and results which others series do not have. Combination of both will highly lead to produce more accurate and reliable results compared to one type of series alone. Panel data analysis is highly recommended if the main purpose of research to estimate relationship at an individual or disaggregated level.

There are several types of panel data models including fixed effect model, random effect model, between estimators, within estimator, dummy variable estimator, first differencing estimator, Feasible Generalized Least Square (FGLS), Ordinary Least Square (OLS), Monte Carlo approaches and many others (Wooldridge, 2012). A lot of research about estimation of panel data has been done. However, most of them just focused on the estimation itself and lack of attention to efficiency and consistency of the estimations.

Estimator is used to infer the value of an unknown parameter in statistical models. This is because in real world the exact value of the population is not known. Reed and Ye (2009) in their research mentioned that the most common estimators in panel data are Generalized Least Square (GLS) and Feasible Generalized Least Square (FGLS). Since variance covariance is often unknown, FGLS is more frequently used rather than GLS. However, the poor performance of the FGLS estimator arises because the true value for variance covariance is unknown (Reed and Ye, 2009). Furthermore, Reed and Ye (2009) also mentioned that Ordinary Least Square (OLS) is also one of the preferable estimators in panel data analysis. Unfortunately, the OLS estimator is generally inconsistent when the independent variable and random error disturbance are correlated. To remedy this inconsistency, one method can be used which is method of Instrumental Variable (IV).

Based in above background, the purpose of the study is to determine the best panel data model to estimate parameters for Sabah construction industries among fixed effects, random effect model and pooled Ordinary Least Squares (OLS). It is important to choose appropriate model which leads a better consistency and efficiency by considering several important aspect such as RMSE and other test if needed in order to choose the best one. The study is limited on the data of Sabah construction Industry which have four cross section variables, which are residential construction, non-residential construction, civil engineering construction and special trade construction.

2. **Literature Reviews**

2.1. Panel data

Panel data refers to data sets consisting of multiple observations on each sampling unit. This could be generated by pooling time-series observations across a variety of cross-sectional units (Baltagi, 2013). An example of panel data are annual unemployment rates of each state over several years, quarterly sales of individual stores over several quarters and wages for the same worker, working at several different jobs. Hsiao (1986) proposed one of the benefits of panel data sets where it provides much larger data sets with more variability and less collinearity among variables compared to typical of cross-section or time series data alone. In addition, he also mentioned other benefits of panel data including panel data sets are more informative and able to control for individuals heterogeneity. Controlling for individual heterogeneity is necessary because it is can cause to bias estimate.

Kasprzyk et al. (1989) said that limitations in panel data sets include problems in the design, data collection and data management of panel surveys. These include the problems of coverage or also known as incomplete account of population of interest, non-response which might be due to the lack of cooperation among respondent or might because of interviewer errors, recall because some of respondent are not remembering correctly, frequency of interviewing, interview spacing, reference period, the use of bounding to prevent the shifting of events from outside the recall period into the recall period and time in sample bias. Another limitation of panel data sets are the distortions due to measurement errors. The measurement errors may arise because of faulty of response due to unclear questions, memory errors, deliberate distortion of responses (e.g., prestige bias), inappropriate informants, miss-recording of responses and interviewer effects.

2.2. Estimations in panel data models

Swamy (1971), Hsiao (1986) and Dielman (1989) said that a simple regression with error components disturbance for one independent variable can be used in the estimation and specification of panel data models. Panel data model with more than one independent variable is written as follows:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \ldots + \beta_k x_{k,it} + u_i + \lambda_t + v_{it} \quad (1)$$

where,
$i = 1,2,\ldots,n$ (i denotes individual, entity),
$t = 1,2,\ldots,T$ (t denotes time),
x_{it} = vector of observations of explanatory variables,
β_k = coefficient of the independent variables,
u_i =an unobserved individual specific effect,
λ_t = an unobserved time specific effect and
v_{it} = zero mean random disturbance with variance, σ_v^2.

Baltagi (1986) stated that if u_i and λ_t denote as fixed parameter to be estimated, this model is known as fixed effect (FE) model. In addition, the FE estimator cannot estimate the effect of any time invariant variable nor can estimate the effect of any individual invariant variable. If u_i and λ_t are random variables with zero means and constant variance σ_u^2 and σ_v^2, this model is known as the random effect (RE) model (Baltagi, 1986). In this model, u_i, λ_t and v_{it} are assumed to be conditionally independent. Baltagi (1986) also mentioned that the RE model can be estimated by using Generalized Least Square (GLS) estimation which can be obtained by using ordinary least square (OLS) regression method. The RE estimator is known by the corresponding GLS estimator of β_k. Note that for this RE model, one can estimate the effects of time invariant and individual invariant variables. Questions about which model perform affectively among RE and FE models often arise. Baillie and Baltagi (1995) derived the asymptotic mean square prediction error for the FE and RE and compared their performance using Monte Carlo simulations.

Beck and Katz (1995) in their research found that, modified version of 'inefficient' OLS for panel data performs substantially better than the asymptotically efficient FGLS estimator in many circumstances. They also said that OLS and FGLS are most common used estimator in panel data sets. Meanwhile, Reed and Ye (2009) proposed the way to compare performance of estimator. In their research, estimator performance was compared on two dimensions. First, root means square error (RMSE) and second is accuracy of estimated confidence intervals. They found that FGLS was the overall best performer on efficiency ground but most worst when it comes to estimating confidence intervals.

In panel data model, asymptotic efficiency concern the limiting value of the variance of any estimator as the sample size increase. In a study conducted by Nickell (1981) found that the least squares dummy variable (LSDV) estimator is not consistent for finite *T* in autoregressive panel data models. Then a number of consistent instrumental variable (IV) and generalized method of moments (GMM) estimators have been proposed in the econometric literature as an alternative to LSDV. Panel data sets are always involved with large number of observations. Arellano and Bond (1991) proposed a GMM estimator which is suitable for panel data with greater number of observation (*n*). Under the random effects model, GLS based on the true variance components is BLUE, and all the feasible GLS estimators considered are asymptotically efficient as *n* and *t* approaches to infinity.

2.3. Applications of panel data models

Applications of panel data in industries have been grown well lately. Bauer et al. (2004) conducted a study on the predictability of stock return in a panel data of individual stocks and perform misspecification test related to the

cross industry heterogeneity. Econometric model that they used can deal with unbalanced panel data, cross sectional correlation among prediction errors and industry specific time effect. Panel data models are also useful in electricity distribution sector. Farsi et al. (2005) in their research applied panel data models in order to find the efficiency of the electricity distribution sector. They compared the estimated coefficients and efficiency of scores across three different panel data models. They used GLS, MLE and RE models. The results of their research indicated that the RE model could be used to measure the possible impacts of unobserved factors such as network effects on efficiency of estimates.

Mariel et al. (2006) estimated parameters of demand equation by applying different types of statistical methodologies using panel data from German car industry. They focused on advertising variables. The important conclusion from their project paper was advertising play an important role but the effectiveness depends on message. Allegretto et al. (2011) did a research about accounting for heterogeneity and selectivity in state panel data. They found that heterogeneity in employment patterns and selectivity among states constituted significant concerns for conventional minimum wage studies. Meanwhile, a research about ownership structure and corporate governance on bank efficiency in the Ghanaian banking industry was conducted by Bokpin (2011). He applied both accounting data and efficiency measures from the period 1999-2007 through panel data analysis on his research. He described that efficiency is found by measuring the differences between the stochastic frontier of estimated translog cost and profit functions.

3. Data and methodology

3.1. Data

The data set was obtained from economic census of construction industry of Sabah in the period of 1992-2010. The data is available in the book Quarterly Construction Statistics which is published by the Department of Statistics Malaysia (DSM, 2014). The main objective of the census was to gather information regarding growth, contribution, composition and distribution of output, employment and other variables related to construction industry in Malaysia especially in Sabah and to assist the Malaysian government in economic planning and formulating policies.

Private sector and individuals can also use the data for economic analysis purposes. The census was primarily done by mail enquiry using a standard questionnaire. The respondents for the census were construction industries in Sabah which were chosen based on type of constructions and number of employee. They were given period of time which was about one month to answer all question completely and return the questionnaires to the Department of Statistics Malaysia. Variables and their definitions in the census of Construction Industry of Sabah in the period of 1992-2010 were value of gross output (OUTPUT), number of establishment per year (ESTABLISHMENT), value of intermediate input (INPUT), value of fixed assets (ASSET), salaries and paid wages (SALARY), and amount of employment (EMPOY).

3.2. Research methodology

Estimated Pooled OLS equation was written in a form similar to the simple regression equation. The method of Pooled OLS estimates was to minimize the sum of squared residuals. The estimated of parameters were chosen simultaneously to make sum of square residuals as small as possible (Wooldridge, 2012). The estimated Pooled OLS regression is written as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \ldots + \hat{\beta}_k x_k \,, \tag{2}$$

where $\hat{\beta}_0$ is the estimate of constant, and $\hat{\beta}_i$ are the estimate of slopes correspond to each explanatory variable.

Since panel data is a combination of cross section and time series data, then it may have cross sectional effects, time effects or both. The effects are either fixed effect or random effect. In fixed effects model, it is desirable to assume difference in intercepts across cross sectional or time series, while in random effect model is more to explore about the difference in error variances. In fixed effect model, there are two ways to do estimations which are within

effect and between effect estimation. The estimators produce identical slope of non-dummy independent variables but they produce different parameter estimates (Wooldridge, 2012). Between method is divided into two, namely between times and between group estimators. Between groups estimations for data of Sabah constructions industries cannot be done in SAS programming since the number of cross section variables is less than number of regressors. While within estimator is supported by PROC PANEL under FIXONE/FIXTWO options in SAS.

Random effects model studies how cross section and/or time series affect the error variance. The random effect model is suitable for *n* individuals (cross sectional units) which are drawn randomly from a large population. In order to estimates random effect model, there are two available estimators. The first one is FGLS method which is used to estimate the variance structure when variance covariance matrix is not known while GLS method is generally used when variance covariance matrix is known. The FGLS estimator is supported by PROC PANEL in SAS under RANONE/RANTWO options. In this paper, FGLS was used since the variance covariance matrix was unknown.

Both fixed and random effect models have one-way and two-way analysis. One-way analysis includes only cross sectional variables in the output while two-way analysis considers both cross sectional and time series variables. Equation for both one-way and two-way for both fixed and random effect models are displayed in Table 1.

Table 1: Fixed and random effect panel data models

Terms	Fixed effect model	Random effect model
Equation	One-way: $y_{it} = (\alpha + u_i) + x_{it}\beta + v_{it}$ Two-way: $y_{it} = (\alpha + u_i + \lambda_t) + x_{it}\beta + v_{it}$	One-way: $y_{it} = \alpha + x_{it}\beta + (u_i + v_{it})$ Two-way: $y_{it} = \alpha + x_{it}\beta + (u_i + \lambda_t + v_{it})$
Intercept	Varying across cross sectional/time series	Constant
Error variance	Constant	Varying across cross sectional/time series
Slope	Constant	Constant
Estimation	Between, Within	FGLS, GLS

where:

y_{it} = dependent variables,

x_{it} = independent variables,

v_{it} = zero mean random disturbance with variance , σ_v^2 ,

u_i = an unobserved individual specific effect,

λ_t = an unobserved time specific effect, and

β = model coefficients.

By referring to the Table 1, fixed effect models treat differences of individual specific effect, u_i , in intercepts and it assume same slope and constant variances across cross sectionals. Since individual specific effect is time invariant, u_i are allowed to be correlated with other independent variables (Wooldridge, 2009). While random effect models assume intercept and slope as a constant. The random effect models treat differences of individual specific effect in error variance.

3.3. Analyzing panel data model is SAS

PROC PANEL was used for estimating parameters in the panel data models. In SAS programming, PROC PANEL provides several options such as FIXONE, FIXTWO, RANONE, RANTWO, BTWNT, BTWNG, and so on. FIXONE and FIXTWO are default for within method. The RANONE and RANTWO are default for FGLS method. In SAS software, both PROC TSCREG and PROC PANEL can be used to handle panel data. Both

procedures produce considerably same results. One-way and two-way random effect and fixed effect model can be done in both PROC PANEL and PROC TSCREG. PROC PANEL is able to deal with both balanced and unbalanced panel data sets. Unfortunately PROC TSCREG is only capable to handle balanced data. Other comparison is PROC PANEL has an option BP and BP2 to conduct the Breucsh Pagan test which can help us to choose between fixed effect and random effect test, but PROC TSCREG does not.

4. Results and Discussion

The models were fit based on balanced panel data of 4 types of constructions over a 12 year period from 1992 to 2010 in Sabah. The sample includes 48 observations with 12 observations per type of instruction. In this paper, the relationship between variable OUTPUT and independent variables was investigated using panel data models. The estimated parameters were listed in Table 2. Most estimated coefficients have signs which were reasonable with economic intuition. Variables INPUT and SALARY have positive effects on the response variable and were showing highly significant since the p values were very small for all types of estimator. Furthermore, the results for coefficient and p value of these two variables were about similar across all estimators. However, estimated parameters for ESTABLISHMENT and EMPLOYEES have larger effects on the OUTPUT compared to other variables. While variable ASSET was having negative effect on OUTPUT for all models except pooled OLS, which suggest that as utilization of ASSET increase, the OUTPUT decreases.

Table 2: Estimated parameters of for modeling Sabah constructions industry

	Pooled OLS		Fixed effect One-way		Fixed effect Two-way		Random effect one-way		Random effect two-way	
	Coeff	p	Coeff	p	Coeff	p	Coeff	p	Coeff	p
Estab	104350	0.095	13826	0.840	64203	0.649	24000	0.717	82337	0.290
Input	1.161	<0.0001	1.144	<0.0001	1.114	<0.0001	1.145	<0.0001	1.145	<0.0001
Employ	4327	0.020	5479	0.0104	5125	0.107	5405	0.001	4679	0.037
Salary	0.698	0.003	0.834	<0.0001	0.835	0.000	0.816	<0.0001	0.804	<0.0001
Asset	0.147	0.070	-0.039	0.618	-0.077	0.418	-0.014	0.849	-0.012	0.876

By definition, MSE of an estimator to measures the average of the difference between the estimator and what is estimated. Small MSE values are needed in statistics because it is closer to actual data and lead to a better estimator. In this study, best model would be chosen based on smallest value of RMSE. Table 3 showed RMSE for each estimator and it was found that the smallest RMSE was obtained from the fixed effect of within method for two-way analysis. The next smallest RMSE was obtained from the random effect of FGLS method for two-way analysis. In both models, two-way analysis provide us better RMSE which is reasonable since the panel data concerns about cross section and time series at the same time. Therefore, in this panel data set the two-way analysis of fixed effect was the efficient estimator.

In the fixed effect two-way model for the Sabah construction industries, at significance level, alpha=0.10, except for pooled OLS, ASSET and ESTABLISHMENT were the only non-significant variables. The reason was that the numbers for establishments were not consistent over the years. In 1994, 1996 and 1998 the numbers of establishments decreased for all type of constructions. This was because of the Malaysian economic crisis at that time which result in changing in economic environmental. This change led the number of establishments in construction industry of Sabah to adopt different modes of operation with regards to the ways they handled risk.

Table 3: RMSE of final reduced model for Sabah construction industry

Models	RMSE
Pooled OLS	21756643
Random effect FGLS method (one-way)	17146596
Random effect FGLS method (two-way)	16469471
Fixed effect within method (one-way)	17158878
Fixed effect within method (two-way)	16400112

In 2010, number of establishment increased dramatically. This was happened in Sabah because on that time Sabah was in the middle of construction of the biggest mall namely 1Borneo Hypermall which needed all types of construction to take part. Any changing in numbers of establishment would affect other variables. Total EMPLOYMENT, INPUT and SALARY affects variable ASSET. Thus, it seemed that establishment cut their labor cost (salaries & wages paid) before they face credit repayment problems. Meanwhile, variable ASSET was not significant so that the effect of the ASSET on the OUTPUT can be ignored even the effect was negative. The negative effect of ASSET indicated that the construction industries in Sabah have debt.

5. Conclusion

The main contribution of this research was the use of panel data model to estimated parameters in Sabah construction industries by using fixed effect model, random effect model and Pooled OLS. In addition, this research presented the estimator of panel data and studied the consistency and efficiency of the estimator by using real data from Statistics Department of Sabah. It was found a more efficient method among pooled OLS, fixed effect and random effect models. It was also found that fixed effect model with two-way analysis is the efficient estimator than others.

References

Arellano, M. and S. Bond., 1991. Some Tests of Specification for Panel Data: Monte Carlo Evidence and An Application to Employment Equations, Review of Economic Studies, 58, 277-97.

Allegretto, S., Dube, A., and Reich, M: Do Minimum Wages Really Reduce Teen Employment? Accounting For Heterogeneity and Selectivity in State Panel Data. IRLE Working Paper. 35:21-28. 2011.

Baillie, R. and B.H. Baltagi:. Prediction from the Regression Model with One-way Error Components, Working paper, Department of Economics, Texas A&M University, College Station, Texas. 1995.

Balestra, P., and M., Nerlove., 1966. Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas, Econometrica, 34, 585-612.

Baltagi, B.H., 1986. Pooling Cross-Sections with Unequal Time-Series Lengths, Economics Letters, 18, 133-136.

Baltagi, B.H., 2013. Econometric Analysis of Panel Data, 5th edition, Chichester, Wiley.

Bauer, R., Pavlov, B., and Schotman, P. C.: Panel Data Models For Stock Returns: The Importance of Industries, University of Maastricht, Report No WP03, Vol. 008. 2004.

Beck, N. and Katz, J., N, 1995. What to do (and not to do) with time series and cross section data, American political science review, 89, 634-639.

Bokpin, G.,A., 2011. Ownership Structure, Corporate Governance and Bank Efficiency: An Empirical Analysis of Panel Data from the Banking Industry in Ghana. Corporate Governance. 13(3):274-287.

Department of Statistics Malaysia, (DSM): Quarterly Construction Statistics 2012. http://www.statistics.gov.my/portal/ download Construction/ files/ Construction/QCS_Q1_2012.pdf

Dielman, T.E., 1989. Pooled cross-sectional and time series data analysis, New York: Marcel Dekker.

Farsi, M., Filippini, M., and Greene, W,: Application of Panel Data Models in Benchmarking Analysis of the Electricity Distribution Sector. CEPE Working Paper. 39. 2005.

Hoch, I., 1962. Estimation of Production Function Parameters Combining Time Series and Cross-Section Data, Econometrica, 30, 34-53.

Hsiao, C., 1986. Analysis of panel data, Cambridge: Cambridge University Press.

Kasprzyk, D., G.J. Duncan, G. Kalton and M.P. Singh, 1989. Panel surveys, New York: John Wiley.

Mariel, P., Lopez, C., and Fernandez, K., 2006. Sales-Advertising Relationship: An Application of Panel Data from The German Automobile Industry. Prague Economic Papers. 12, 22-28.

Nickel, S., 1981. Biases in Dynamics Models with Fixed Effect, Econometrica, 49,1417-1246.
Reed, W.R., and Ye, H., 2009. Which panel data estimator should I use? Applied economic. 43 (8), 985-1000.
Swamy, P.A.V.B., 1971. Statistical inference in random coefficient regression models, New York: Springer- Verlag.
Wooldridge, J. M., 2012. Introductory Econometrics a Modern Approach: 5th edition, Michigan: Cengage Learning.